Journal of Biomedical Informatics 43 (2010) 569–577

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Exploring the ncRNA–ncRNA patterns based on bridging rules

Feng Chen [a], Yi-Ping Phoebe Chen [a,b,*]

[a] Faculty of Science, Technology and Engineering, La Trobe University, Bundoora, Vic. 3086, Australia
[b] Australia Research Council Centre of Excellence in Bioinformatics, Australia

ARTICLE INFO

Article history:
Received 8 July 2009
Available online 10 February 2010

Keywords:
ncRNAs
Bridging rules
Entropy
miRNA
Joint entropy
Mutual information

ABSTRACT

ncRNAs play an important role in the regulation of gene expression. However, many of their functions have not yet been fully discovered. There are complicated relationships between ncRNAs in different categories. Finding these relationships can contribute to identify ncRNAs' functions and properties. We extend the association rule to represent the relationship between two ncRNAs. Based on this rule, we can speculate the ncRNA's function when it interacts with other ncRNAs. We propose two measures to explore the relationships between ncRNAs in different categories. Entropy theory is to calculate how close two ncRNAs are. Association rule is to represent the interactions between ncRNAs. We use three datasets from miRBase and RNAdb. Two from miRBase are designed for finding relationships between miRNAs; the other from RNAdb is designed for relationships among miRNA, snoRNA and piRNA. We evaluate our measures from both biological significance and performance perspectives. All the cross-species patterns regarding miRNA that we found are proven correct using miRNAMap 2.0. In addition, we find novel cross-genomes patterns such as (hsa-mir-190b → hsa-mir-153-2). According to the patterns we find, we can (1) explore one ncRNA's function from another with known function and (2) speculate the functions of both of them based on the relationship even we do no understand either of them. Our methods' merits also include: (1) they are suitable for any ncRNA datasets and (2) they are not sensitive to the parameters.

1. Introduction

A non-coding RNA (ncRNA) is a RNA molecule that is not translated into a protein. It plays an important role in regulating gene expression by various mechanisms. Many approaches have been applied to study ncRNA, resulting in the discovery of a large number of new ncRNAs [1]. However, the functions of many ncRNAs are still not clear. Gene regulation networks inspire us to speculate a ncRNA's function based on the relationships between ncRNAs because all the ncRNAs collaborate with other functionally associated genes by interaction and reaction [2–4].

There are many papers concerning the relationships involved with ncRNAs. However most of them have focused on exploring how miRNAs regulate mRNAs [5–10] while ignoring other relationships regarding ncRNAs: the relationships between cross-species or cross-genomes ncRNAs. These relationships can help us determine ncRNA functions and understand the evolutionary process. We have applied bridging rule mining [11] to find the relationships between ncRNAs. A bridging rule is an association rule where its antecedent and action belong to different conceptual clusters. For example, a pattern "$A \rightarrow B$" found by the bridging rule algorithm is not only a bridging rule but also an ncRNA–ncRNA pattern if A belongs to snoRNA and B belongs to miRNA. According to whether the function of the ncRNA is already known or not, an ncRNA–ncRNA can be divided into four situations:

(1) known ncRNA → known ncRNA
(2) unknown ncRNA → unknown ncRNA
(3) unknown ncRNA → known ncRNA
(4) known ncRNA → unknown ncRNA

In situation 1, we know the functions of both sides of a pattern. A ncRNA–ncRNA pattern can help us find potential knowledge in them. As for situation 2, finding a relationship between the antecedent and action will be helpful for us to determine their functions. Situations 3 and 4 can be combined into one category. According to our background knowledge and this pattern, we can determine the function of the unknown ncRNA from another known ncRNA. To find these patterns, there are three difficulties we have to deal with:

(1) How to determine the evaluation standard.
(2) How to analyze the computational results.
(3) How to prove the accuracy from the biological point of view.

* Corresponding author. Address: Faculty of Science, Technology and Engineering, La Trobe University, Bundoora, Vic. 3086, Australia.
E-mail address: phoebe@deakin.edu.au (Y-P.P. Chen).

1532-0464/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2010.02.006

We combine linear standard with non-linear standard for dealing with the difficulties. We use similarity as the linear standard. Entropy, including joint entropy and mutual information, is taken as the non-linear standard. We determine if a pattern is significant based on these two standards. At the same time, we understand the patterns according to genome background and biological literatures.

We have developed two measures, which are joint entropy and mutual information, respectively, to study the relationships in ncRNA–ncRNA patterns from different angles. The relationship in a ncRNA pair belonging to different categories indicates a kind of interaction which is of higher similarity and more information exchange than other ncRNA pairs. First of all, we use the maximum similarity mentioned in example 1 to determine if two ncRNAs are similar. Similarity is a basic measure to guarantee that that the patterns we find are significant biologically. Second of all, we employ joint entropy or mutual information to evaluate how much information in this ncRNA pair. In this step, we consider the ncRNAs that impact this ncRNA pair to make sure our patterns are analyzed in gene regulation networks. Last of all, for ordering our patterns to highlight the most significant pairs, we have introduced a ranking measure to combine these two criteria. We can find significant ncRNA pairs based on the above three steps.

Currently, most research concerning ncRNA focuses on the prediction of secondary structure [12–14], which is very important because many RNAs' functions are determined by their structures. From the evolution angle, RNA functions are preserved better in secondary structures than in primary sequence. However, we focus on the ncRNA patterns not only across the species and genomes but also in different classification criteria. Our methods can provide different patterns when classification criteria are different. When we analyze patterns using other criteria other than species, we cannot determine how the ncRNA structure affects the results. We use sequence based similarity to evaluate the similarity of two ncRNAs. The validation of our cross-species miRNA patterns by miRNAMap [15] proves that our methods can find the related miRNAs even without taking secondary structure into consideration.

We apply our measures to three different datasets to find relationships between ncRNAs, including miRNA, snoRNA and piRNA. MicroRNA (miRNA). One of the small molecular RNA families, is a very small section of non-coding RNA sequence. These regulate several biological processes [16–18,52,53] and are closely related with mRNA and cancer [19,20]. Small nucleolar RNAs (snoRNAs) are an abundant group of non-coding RNAs, which are mainly involved in post-transcriptional modification of rRNAs in *eukaryotes* [21–24]. The RNAs, the length of which is frequently 30 nucleotides, are called Piwi-interacting RNAs (piRNAs). The piRNAs generally distribute across only one genomic strand or distribute on two strands but in a divergent, nonoverlapping manner [25,26]. Some of them are derived from transposons regulating the silence of repetitive elements [27]. Some of them can repress transposons in mammals [28]. It was also assumed that piRNAs would have something to do with sperm generation although their functions are not fully known. All the three classes of ncRNAs are important and unique. We can demonstrate that our methods are useful and significant by analyzing these datasets. The experiment results and the theoretical background can demonstrate that our contributions include:

(1) Design two measures to find interesting ncRNA–ncRNA patterns.
(2) Analyze and evaluate the relationships we find, and then speculate their functions from the biological point of view.

The structure of this paper is as follows: In Section 2, we give a description about the bridging rule and explain the details about how to apply it in our research. Section 3 is our experiments in which we analyzed our measures from the performance and biological significance perspectives. Section 4 includes discussion about our measures, computational results and conclusion.

2. Methods

A bridging rule [11] is an association rule where the antecedent and action come from different conceptual clusters. It includes two sub-algorithms: joint entropy and mutual information. Firstly, a similarity measure is used to determine the similarities between the objects. Each object is composed of a sequence. For example, if A = aabcd and B = aaccc, then the similarity of A and $B = |A \cap B|/|A \cup B| = 3/5 = 0.6$. Secondly, based on these objects of high similarity, joint entropy or mutual information is employed to measure how much information these objects have. The more information they have, the closer they are. In this paper we have developed these two sub-algorithms to find the relationships between ncRNAs. We choose entropy [29] because the relationships between ncRNAs are so complex that pure linear measures cannot evaluate them.

2.1. Joint entropy measure

The joint entropy measures how much information is contained in a joint system of two random variables. In this paper, each ncRNA sequence represents a variable. The structure of joint entropy measure is as follows (see Fig. 1).

Here C_1 and C_2 are two classes in one dataset. g_1 and g_2 are two ncRNAs we want to analyze. $\{g_{11}, \ldots, g_{1n}\}$ ($\{g_{21}, \ldots, g_{2m}\}$) means the nearest neighbors [30] of $g_1(g_2)$. The line between two ncRNAs indicates there is a relationship between them. The shorter line does not mean the closer relationship, and vice versa. When calculating joint entropy, we just consider ncRNAs in these two nearest neighbor sets while ignoring others which are little related with g_1 and g_2 to reduce space- and time–cost. Therefore the first step is to calculate the similarity of two ncRNAs based on the ncRNAs' sequences. We have improved the similarity definition in bridging rules to deal with the ncRNAs with different lengths. Example 1 is about similarity calculation:

Example 1. How to calculate the similarity of two ncRNAs

g_1 = AAGGUU
g_2 = AAAGGUGUAA

Let g_1, g_2 be two ncRNAs. Let us use length(g_1)(length(g_2)) to denote the length of $g_1(g_2)$, sim(g_1, g_2) indicates the similarity of g_1 and g_2. When length(g_1) is not the same as length(g_2), sim(g_1, g_2) means the maximum similarity of g_1 and g_2. For example, in Table 1, sim(g_1, g_2) has five values. We suggest sim(g_1, g_2) is 0.5, which is the biggest in all the possible values. We use this method to save running time and space, instead of aligning the sequences allowing the gaps. The experiments in Section 3 will illustrate this strategy is effective and promising.

rel_sim(g_1, g_2) denote the relative similarity of g_1 and g_2:

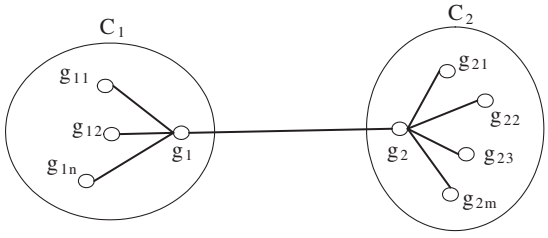

Fig. 1. The structure of joint entropy.

Table 1
The example of how to calculate $\mathrm{sim}(g_1, g_2)$.

×AAGGUU××× AAAGGUGUAA $\mathrm{sim}(g_1, g_2) = 0.5$	××××AAGGUU AAAGGUGUAA $\mathrm{sim}(g_1, g_2) = 0.1$
×××AAGGUU× AAAGGUGUAA $\mathrm{sim}(g_1, g_2) = 0.2$	AAGGUU×××× AAAGGUGUAA $\mathrm{sim}(g_1, g_2) = 0.4$
××AAGGUU×× AAAGGUGUAA $\mathrm{sim}(g_1, g_2) = 0.3$	

$$\mathrm{rel\_sim}(g_1, g_2) = \mathrm{sim}(g_1, g_2) \times \frac{\mathrm{length}(g_1)}{\mathrm{length}(g_2)} = 0.5 \times \frac{6}{10} = 0.3$$

So our formula of $\mathrm{rel\_sim}(g_1, g_2)$ is:

$$\mathrm{rel\_sim}(g_1, g_2) = \begin{cases} \mathrm{sim}(g_1, g_2) \times \frac{\mathrm{length}(g_1)}{\mathrm{length}(g_2)} & \mathrm{length}(g_2) > \mathrm{length}(g_1) \\ \mathrm{sim}(g_1, g_2) \times \frac{\mathrm{length}(g_2)}{\mathrm{length}(g_1)} & \mathrm{length}(g_1) > \mathrm{length}(g_2) \end{cases} \quad (1)$$

In addition, we define:

$$H(g_1) = -\sum_{i=1}^{n} p(g_1, g_{1i}) \log p(g_1, g_{1i}) \quad (2)$$

to calculate the entropy of g_1 and g_{1i} where $p(g_1, g_{1i}) = \frac{\mathrm{rel\_sim}(g_1, g_{1i})}{\sum_{i=1}^{n} \mathrm{rel\_sim}(g_1, g_{1i})}$. The definition of $H(g_2)$ is the same. Then the information exchange of g_1 and g_2, denoted by $H(g_1, g_2)$, is:

$$H(g_1, g_2) = H(g_1) + H(g_2) \quad (3)$$

We define a rel_sim threshold as "min_rel_sim" and an joint entropy as "min_joint". Every pattern which is satisfied with the following two criteria will be considered to be interesting:

$$\begin{cases} \mathrm{rel\_sim}(g_1, g_2) > \mathrm{min\_rel\_sim} \\ H(g_1, g_2) > \mathrm{min\_joint} \end{cases} \quad (4)$$

2.2. Mutual information measure

Here we need two definitions: mutual information ($I(X; Y)$) and conditional mutual information ($I(X; Y \mid Z)$). The first indicates how much information a ncRNA set $X(Y)$ can obtain from another ncRNA set $Y(X)$. The second one measures the amount of mutual information of X and Y given the third ncRNA set Z. So $I(X; Y \mid Z) - I(X; Y)$ indicates how much information will be lost when Z is ignored. Namely, how much information Z has. The structure of the mutual information measure is shown in Fig. 2.

In Fig. 2, C_1 and C_2 are two classes in one dataset. $Z = \{z_1, z_2\}$ means a set of two ncRNAs that we want to analyze, which can be extended into the research of multiple ncRNAs. $X = \{x_1, x_2, \ldots, x_n\}$ ($Y = \{y_1, y_2, \ldots, y_m\}$) indicates the nearest neighbors of $z_1(z_2)$. For example there is a rule (mdo-let-7i → gga-let-7i) in Table 5. The dataset of this rule is "let7". C_1 is class *Monodelphis domestica*, C_2 is class *Gallus gallus*. z_1 and z_2 indicate mdo-let-7i gga-let-7i, respectively.

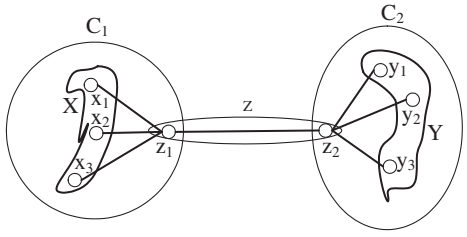

Fig. 2. The structure of mutual information.

We define two thresholds: min_rel_sim for similarity and min_mutual for mutual information. Those patterns which are satisfied with the following two inequations will be considered to be significant and noticeable.

$$\begin{cases} \mathrm{rel\_sim}(z_1, z_2) > \mathrm{min\_rel\_sim} \\ |I(X; Y|Z) - I(X; Y)| > \mathrm{min\_mutual} \end{cases} \quad (5)$$

where $I(X; Y) = \sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i)p(y_j)}$ and $I(X; Y|Z) = \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{2} p(x_i, y_j, z_k) \log \frac{p(x_i, y_j|z_k)}{p(x_i|z_k)p(y_j|z_k)}$. The definitions of $p(x_i)$ and $p(x_i, y_j)$ are as follows:

$$p(x_i, y_j) = \frac{\mathrm{rel\_sim}(x_i, y_j)}{\sum_{i'=1}^{n} \sum_{j'=1}^{m} \mathrm{rel\_sim}(x_{i'}, y_{j'})} \quad (6)$$

$$p(x_i) = \frac{\sum_{j=1}^{m} \mathrm{rel\_sim}(x_i, y_j)}{\sum_{i'=1}^{n} \sum_{j'=1}^{m} \mathrm{rel\_sim}(x_{i'}, y_{j'})} \quad (7)$$

The calculation of $p(x_i, y_j, z_k)$ is similar. The calculation of $p(x_i|z_k)$ can be found through the formula $p(x_i|z_k) = \frac{p(x_i, z_k)}{p(z_k)}$.

2.3. Ranking measure

The patterns from these two algorithms have two values: rel_sim and joint entropy value (or mutual information value). From Table 2 we can see that a pattern might not have a high rel_sim even if it has high joint entropy value (mutual information value). We propose a ranking measure to combine these two values to identify any significant patterns, the formula of which is shown in the following formula.

$$\mathrm{RankAll}(\mathrm{P}^*) = \frac{1}{\left(\frac{\mathrm{RankSim}(\mathrm{P}^*)}{\mathrm{SumNum}}\right)^{\alpha} + \left(\frac{\mathrm{RankEntro}(\mathrm{P}^*)}{\mathrm{SumNum}}\right)^{\beta}} \quad (8)$$

Before employing formula 8, we list all the patterns according to the decreasing order of similarity and of joint entropy value (mutual information value), respectively. In formula 8, P* represents a pattern we are considering. SumNum is the number of patterns we obtain from our algorithms. RankSim(P*) indicates the position of P* in similarity list. RankEntro(P*) indicates the position of P* in joint entropy value (mutual information value) list. α and β are two coefficient factors larger than 1. When $\alpha > \beta$, we consider more similarity than joint entropy value (mutual information value). When $\alpha < \beta$, we think that entropy value is more important than similarity. In example 2 and the following experiments, we let both of them be 1 for simplifying. RankAll(P*) represents a position of P* taking similarity and joint entropy value (mutual information value) into consideration.

Table 2
Pattern list.

Patterns	RankSim	RankEntro
P^*_1	1	5
P^*_2	2	3
P^*_3	3	1
P^*_4	4	4
P^*_5	5	6
P^*_6	6	10
P^*_7	7	2
P^*_8	8	7
P^*_9	9	8
P^*_{10}	10	9

Example 2. In Table 2, we list 10 patterns by the decreasing order of the similarity. We can see a pattern with high similarity might not have a high entropy value, such as P^*_1. Let us take P^*_1, P^*_2 and P^*_3 as an example:

$$\mathrm{RankAll}(P^*_1) = \frac{1}{\left(\frac{\mathrm{RankSim}(P^*_1)}{\mathrm{SumNum}}\right)^{\alpha} + \left(\frac{\mathrm{RankEntro}(P^*_1)}{\mathrm{SumNum}}\right)^{\beta}} = \frac{1}{\left(\frac{1}{10}\right)^1 + \left(\frac{5}{10}\right)^1} = 1.67$$

$$\mathrm{RankAll}(P^*_2) = \frac{1}{\frac{2}{10} + \frac{3}{10}} = 2$$

$$\mathrm{RankAll}(P^*_3) = \frac{1}{\frac{3}{10} + \frac{1}{10}} = 2.5$$

So $\mathrm{RankAll}(P^*_3) > \mathrm{RankAll}(P^*_2) > \mathrm{RankAll}(P^*_1)$, which indicates P^*_3 is more important than the other two patterns when we take both of the parameters into account.

3. Results

3.1. Dataset preparation

We have conducted our experiments by using real datasets and synthesized datasets. The real datasets are obtained from miRBase [31] and RNAdb, which are used to find interesting patterns. The miRNAMap release 2.0 (http://mirnamap.mbc.nctu.edu.tw/index.php) [15] is used to verify that our miRNA patterns with respect to miRNA are correct and reasonable because it collects known information of miRNAs, which includes the secondary structure and related miRNAs by comparative sequence analysis [42]. The synthesized datasets are for analyzing the influence of data features, such as class number and the average length of the sequence, because they are hard to be explored in real datasets. The analysis of real datasets is used to prove that our measures are effective and promising. The comparison between synthesized datasets is to illustrate that our measures are not sensitive to the parameters and suitable for any datasets.

3.1.1. Real datasets preparation

We use the dataset of hairpin.fa consisting of FASTA format sequences of all miRNA hairpins, chosen from miRBase. We choose hairpin miRNA [32–34] instead of mature miRNA because (1) it is the first stage of generating mature miRNA so that we can find more basic relationships between miRNAs and (2) the length of hairpin miRNA is much longer than that of mature miRNA (about ~70 nt vs. ~22 nt), which is easier to identify which miRNAs are different because they include more differences than mature miRNAs.

Our first step is to extract data from hairpin.fa to construct two datasets. One, named "let7", consists of all miRNAs belonging to the let-7 family but in different species [35–39]. For example, mdo-let-7i and gga-let-7i belong to let-7 family. But the first one belongs to *M. domestica*. The second one is one member of *G. gallus*. The let-7 miRNAs regulate developmental timing in *C. elegans*. They are an important paradigm for investigations of miRNA function in mammalian development. The other dataset, called "hsa", consists of diverse miRNAs which belong to human but occur in different genomes. For instance, hsa-mir-92b belongs to mir-25 family. hsa-let-7f-1 is one of let-7 family. But both of them belong to *Homo sapiens*.

From RNAdb, We select 3 RNA classes which are microRNA, piRNA, snoRNA. These three classes constituted the third dataset which we call "rna". The details of these three datasets are shown in Table 3.

Table 3
The real datasets summary.

Dataset	Class number	Size	Description
let7	10	114	All genes in let-7 but in different spaces
hsa	14	290	All genes in human but in different genomes
rna	3	150	microRNA, snoRNA, piRNA

3.1.2. Synthesized datasets preparation

Besides similarity and entropy which we can regulate to mine different patterns, there are still two features, which are different in each dataset: the number of class and the average sequence length. For example, "let7" has 10 classes. "rna" includes three classes. In "rna", piRNA is much shorter than snoRNA. It is hard to evaluate exactly how one feature influences our measures in real datasets. However, the analysis of these features can help us understand the advantages and disadvantages of our measures. We construct two datasets using the IBM data generator (IBM) [40] which generate transaction data for mining association rules [41,54]. Here we consider every transaction to be a ncRNA and one item in the transaction to be a base. Each dataset we construct has four sub-datasets. The one named "classnumber" consists of four sub-datasets with different numbers of classes. The one named "length" is composed of four sub-databases with different average sequence lengths. Table 4 summaries the details about these two datasets.

3.2. Biological significance of patterns analysis

For demonstrating that our patterns are significant biologically, we choose the first seven patterns from "let7", the first five patterns from "hsa" and the first five patterns from "rna". We list all of them in Table 5. In this table, the first seven patterns from "let7" can be used to evaluate the relationships between miRNAs belonging to let-7 but in different species such as *M. domestica*, *G. gallus*, *H.* sapiens and so on. The next five from "hsa" patterns include miRNAs in *H. sapiens*. In "rna", we classify the ncRNAs as miRNA, snoRNA and piRNA. So the last five patterns indicate there are relationships between ncRNAs belonging to different categories.

Compared to other patterns, all of these patterns are of high RankAll values. They are extracted to show that our measures can find interesting relationships because it is impossible to list all the patterns. We use miRNAMap to identify that our patterns 1–7 are reasonable. Through enquiring miRNAMap, all the relationships in the first seven patterns can be proven correct. Patterns 1–3 are related with mdo-let-7i. Its relationships with gga-let-7i, hsa-let-7f-1 and rno-let-7f-1 can be identified in its related gene list [43,44]. The relationships in patterns 4–7 can be also found too [43,45,46]. In addition, there are always so many related genes for one particular gene in miRNAMap. For example, mdo-let-7i has 130 related genes. However using our joint entropy measure, we find 19 genes which are related with mdo-let-7i. As for pattern 4, there are 130 genes related with mdo-let-7d in miRNAMap. In our study we only found 9 in our analysis result. These two examples illustrate that we can reduce the size dramatically of the related gene list because we take entropy theory into consideration. Therefore we can find more significant gene patterns which are ignored before.

Table 4
The synthesized datasets summary.

Dataset name	Class number	Average length	Size
Classnumber	3, 4, 5, 6	7	127
Length	3	3, 5, 7, 10	85

Table 5
The patterns we choose from our experiment results. In description column, the first line in very grid describes the antecedent of the pattern. The second line is for the action. rel_sim indicates the similarity of two genes. Joint indicates the value of joint entropy of two genes. Mutual is the value of mutual information of two genes. In "let7", min_rel_sim = 0.5, min_joint = 3, min_mutual = 0.003, k_num (the number of nearest neighbors) = 5. When joint entropy measure is employed, we find 539 patterns. When mutual information measure is employed, we find 623 patterns. In "hsa", min_rel_sim = 0.3, min_joint = 2, min_mutual = 0.0005, k_num = 7. We find 127 patterns using joint entropy measure and 342 patterns using mutual information measure. In "rna", min_rel_sim = 0.25, min_joint = 2.5, min_mutual = 0.0005, k_num=. Joint entropy measure can find 171 patterns and mutual information measure can find 246 patterns. The direction of the pattern is not very important. A pattern just means a set of two genes. There is no difference between "$A \rightarrow B$" and "$B \rightarrow A$".

ID	Dataset	Pattern	rel_sim	Joint	Mutual	Description
1	let7	mdo-let-7i → gga-let-7i	1	3.184	1.409	*Monodelphis domestica* let-7i stem-loop *Gallus gallus* let-7i stem-loop
2	let7	mdo-let-7i → hsa-let-7f-1	0.591	3.18	0.0134	*Monodelphis domestica* let-7i stem-loop *Homo sapiens* let-7f-1 stem-loop
3	let7	mdo-let-7i → rno-let-7f-1	0.573	3.195	0.017	*Monodelphis domestica* let-7i stem-loop *Rattus norvegicus* let-7f-1 stem-loop
4	let7	mdo-let-7d → hsa-let-7d	0.801	3.213	0.00245	*Monodelphis domestica* let-7d stem-loop *Homo sapiens* let-7d stem-loop
5	let7	gga-let-7a-1 → mmu-let-7f-1	0.556	3.205	0.0029	*Gallus gallus* let-7a-1 stem-loop *Mus musculus* let-7f-1 stem-loop
6	let7	gga-let-7d → hsa-let-7b	0.592	3.205	0.00248	*Gallus gallus* let-7a-1 stem-loop *Homo sapiens* let-7d stem-loop
7	let7	tni-let-7g → xtr-let-7f	0.843	3.209	0.00506	*Tetraodon nigroviridis* let-7g stem-loop *Xenopus tropicalis* let-7f stem-loop
8	hsa	hsa-mir-892b → hsa-mir-320	0.378	3.117	0.00526	*Homo sapiens* miR-892b stem-loop *Homo sapiens* miR-320 stem-loop
9	hsa	hsa-mir-7-3 → hsa-let-7f-1	0.354	3.209	0.00143	*Homo sapiens* miR-7-3 stem-loop *Homo sapiens* let-7f-1 stem-loop
10	hsa	hsa-mir-871 → hsa-mir-197	0.376	3.178	0.00768	*Homo sapiens* miR-871 stem-loop *Homo sapiens* miR-197 stem-loop
11	hsa	hsa-mir-190b → hsa-mir-153-2	0.379	3.154	0.00194	*Homo sapiens* miR-190b stem-loop *Homo sapiens* miR-153-2 stem-loop
12	hsa	hsa-mir-92b → hsa-let-7f-1	0.354	3.205	0.00411	*Homo sapiens* miR-92b stem-loop *Homo sapiens* let-7f-1 stem-loop
13	rna	hsa-mir-425 → Mmu_MBII-115	0.306	4.496	0.00204	*Homo sapiens* miR-425 stem-loop *Mus musculus* clone MBII-115 C/D box snoRNA, partial sequence
14	rna	SNO1004 → MIR1460	0.315	4.599	0.00124	*Mus musculus* clone MBII-115 C/D box snoRNA, partial sequence *Gorilla gorilla* miR-101 stem-loop
15	rna	SNO1034 → PIR10001	0.31	4.555	0.0097	*Mus musculus* clone MBII-426 C/D box snoRNA, partial sequence *Mus musculus* piRNA piR-17002, complete sequence
16	rna	PIR10007 → SNO1027	0.356	4.588	None	*Mus musculus* piRNA piR-17008 *Mus musculus* clone MBII-13 C/D box snoRNA,
17	rna	SNO1043 → PIR10031	0.325	None	0.00088	*Mus musculus* clone MBI-15 H/ACA box snoRNA *Mus musculus* piRNA piR-17032

Specifically, the first pattern indicates a relationship between *M.* domestica and *G.* gallus [43]. Pattern 5 represents the relationship between *Mus musculus* and *G.* gallus [43,45,47–49]. *M.* domestica is a gray short-tailed opossum. *M.* musculus is a house mouse. *G.* gallus is the red junglefowl which is a tropical member of the pheasant family. It is the direct ancestor of the domestic chicken. *M.* domestica and *M.* musculus are both mammals but *G.* gallus is not. So compared with the relationship between *M.* domestica and *G.* gallus or between *M.* musculus and *G. gallus*, there is closer relationship between *M.* domestica and *M.* musculus from the point of view of family let-7.

We extend pattern analysis from cross-species to cross-genomes to explore novel patterns. For example pattern 8 indicates a relationship between hsa-mir-892b and hsa-mir-320. According to [60], hsa-mir-320 is related with prostate cancer [56]. By EMBL-EBI (http://www.ebi.ac.uk/), hsa-mir-892b can hit gene SLC43A1, which up-expresses in prostate cancer. This kind of "indirect" relationship cannot be explored by other algorithms. The relationship in pattern 11 can be identified by the report 2009 from Biosietta (http://www.biosettia.com/). This relationship indicates that hsa-mir-190b [49] and hsa-mir-153-2 [45,49] are both related with lentiviral viruses. We also analyze pattern 12 (hsa-mir-92b → hsa-let-7f-1) in miRNAMap [49,50] hsa-mir-92b has 62 related miRNAs and hsa-let-7f-1 has 129. But they are not included in each other's related miRNA list because this database only has cross-species related miRNAs. Because the mature sequence of hsa-mir-92b [5,49,50] represents the most commonly cloned form from large-scale cloning studies, we propose hsa-let-7f-1 [45,46,49–51] might have this feature too. In addition, we also extend our methods to pattern analysis among miRNA, snoRNA and piRNA. We speculate there are relationships between them according to the last five patterns.

From Table 5, we notice that the results of the joint entropy measure are very different from those of the mutual information measure because they focus on different aspects. For example, in "hsa", joint entropy of pattern 9 is 3.209 and mutual information of it is 0.00143. In pattern 8, joint entropy is 3.178 and mutual information is 0.00768. Moreover, there are patterns which are only found by one measure such as pattern 14 and pattern 15. That means a pattern with big joint entropy might not have big mutual information. This is because in joint entropy measure, we think a pattern is important if it has enough information compared to other patterns. In mutual information measure, we consider the two ncRNAs of a pattern and their nearest neighbors to be a set. We calculate the information amount change in this set. The more the information changes, the more important the pattern is. Therefore we do not compare these two methods with each other.

3.3. Datasets features analysis

The datasets with different ncRNAs have different features, which will affect our results. We list five features, including cluster number, average ncRNA sequence length, the number of the nearest neighbor, joint entropy and mutual information. We use two

criteria, which are the running time and pattern number, to prove that our measures are suitable for any kind of ncRNA datasets and not sensitive to the parameters. All the computational results are obtained on Windows XP, 2.0 GHz CPU, 1.0 Gb memory and Perl. The time unit involved in the experiments is second. In the following four figures, Fig. 3 is about the first two features while Figs. 4–6 focus on the other three.

For comparing with three datasets and finding mutual trends within them, we normalize all the results from our experiments and show them in Figs. 4–6. In Fig. 4, with higher similarity, pattern number and running time become fewer very sharply from the beginning. In Fig. 5, the number of the nearest neighbors is not related to pattern number. In Fig. 6, joint entropy value cannot change pattern number firstly. But then, pattern number is influenced deeply by joint entropy. It decreases dramatically when joint entropy value becomes bigger. To sum up, the tendencies of three databases are very similar. The number of patterns decreases when the parameters increase although the decreasing speeds are different. This indicates higher parameters will provide us fewer results. But lower parameters are useful for comprehensive analysis. At the same time, the order of results does not change too much. Important patterns are always located at the first part of the result list. The inconsistency in running time does not influence the results because they are all practical. From these three figures, we can demonstrate our algorithms are not sensitive to the parameters and dataset types.

Let us take Fig. 7 as an example to explain mutual information measure. In this figure, the pink line represents mutual information. The blue line represents similarities of patterns. The centre of the circle and three con concentric circles divide possible value scope into three intervals: 0–0.5, 0.5–1, 1–1.5 and 1.5–2. The numbers around the outside circle indicates that the circle is divided averagely into 20 parts. Each radius (visible and invisible) indicates a pattern. Its mutual information and similarity are marked on this radius. We use this figure to explore the relationships between similarity and mutual information. We can see the relation be-

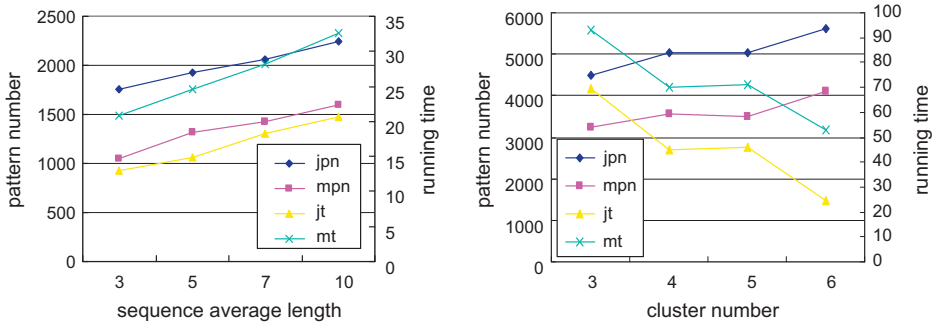

Fig. 3. The analysis of sequence average length and cluster number. The left one (the right one) shows how the average sequence length (cluster number) influences the running time and pattern number. jpn (mpn) means the number of patterns we find by joint entropy measure (mutual information measure). jt (mt) represents the running time we need when joint entropy measure (mutual information measure) is used. In the left graph, the running time and the number of patterns increases as the average sequence length becomes longer. But in the right one, both of them decrease with the increase of the number of clusters.

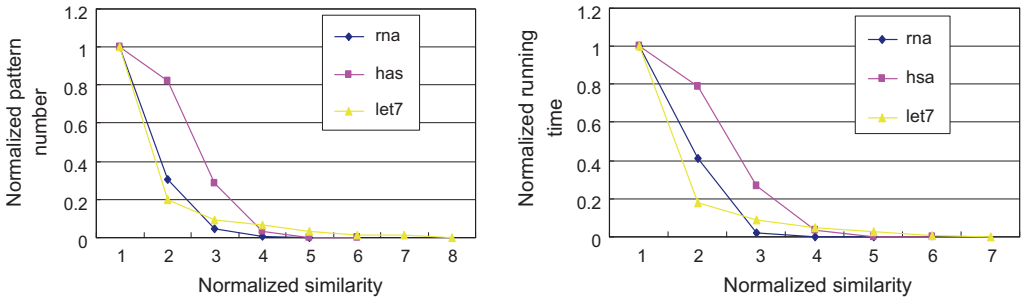

Fig. 4. The normalized tendencies of similarity in three datasets. The horizontal axis indicates normalized similarity. The vertical axis of the left one represents normalized pattern number. The vertical axis of the right one is running time.

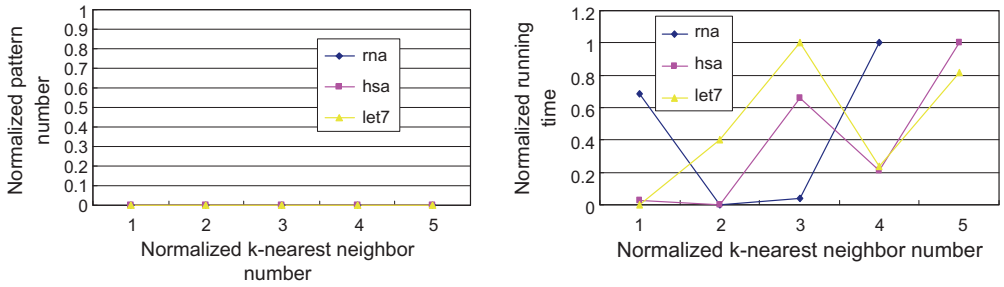

Fig. 5. The normalized tendencies of the number of the nearest neighbor in three datasets. The horizontal axis represents the normalized nearest neighbor number. The vertical axis of the left one represents the normalized pattern number. The vertical axis of the right one represents the normalized running time.

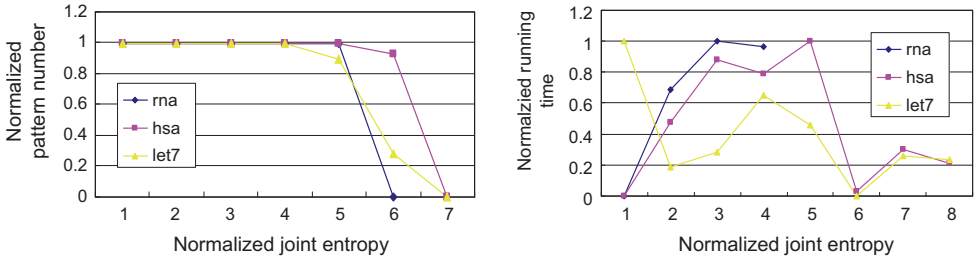

Fig. 6. The normalized tendencies of joint entropy values of three datasets. The horizontal axis represents the normalized joint value. The vertical axis of the left one means the normalized pattern number. The vertical axis of the right one means about the normalized running time.

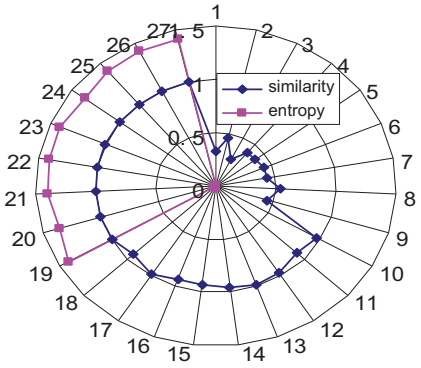

Fig. 7. The relationship between mutual information and similarity in dataset let7.

tween these two parameters can be divided into three parts. When similarity is 1, the value of mutual information is more than 1, between 1.3 and 1.48. When similarity is slightly less than 1, mutual information becomes very small dramatically within the range of 0.1 or 0.2, Mutual information becomes too small to be ignored when similarity is less than 0.5. So the mutual information measure is very difficult to be influenced by parameters because all of the results can be divided into three parts. It indicates that this method is not sensitive to the parameters.

4. Discussion and conclusion

ncRNAs have been explored from distinct angles [61–65]. For example, ncRNA prediction always takes advantage of secondary structure and comparative analysis for the higher accuracy [57–59]. In addition, the exploration ncRNA also focuses on classifying ncRNAs, determining the annotation of known ncRNAs and identifying unknown ncRNAs [63–65]. The relationships between miRNAs and mRNAs take much attention because the regulation between them has been proven to be related with distinct diseases such as cancer [8–10,16,17,19]. However, biologists are still unclear on many ncRNA functions. At the same time, very few materials are concerned the relationships between ncRNAs. In this paper, we have proposed two measures based on bridging rule to find significant relationships between ncRNAs, which can explored the ncRNA functions. These relationships are very different from those from the traditional sequence clustering because the ncRNAs in the relationship come from different genomes. The traditional sequence clustering is good at finding the similar genes in the same genome. Compared with that, our methods can find the cross-genome relationships which contribute to understand the gene regulation mechanism.

Bridging rule is a kind of association rule which is specially designed for finding relationships between objects belonging to different conceptual clusters based on information theory. We apply it to our research. In every dataset that we employ, all the genes are divided into different clusters. But there are biological relationships between them. For example, in “let7”, all of them belong to gene family let-7 but in different species. These datasets makes sure that we can obtain significant relationships based on biological background.

Entropy theory has already been applied to the research of ncRNAs [55,62–65]. Shannon base-pair entropy measure [63,64] is employed to predict ncRNA secondary structures from sequences. Mutual information and relative entropy are taken to align RNAs [62]. Entropy theory [65] is also used to represent and evaluate the secondary structures in pseudoknots. These literatures perform better than the similar literature because they have introduced non-linear feature into the complicated ncRNA exploration. In this paper, we have proposed two measures to find these relationships between ncRNAs in distinct clusters, one of which is based on joint entropy and the other is based on mutual information. The first measure has two thresholds: relative similarity and joint entropy. If a gene pair has higher similarity and bigger joint entropy, then it is more interesting. The second measure has also two thresholds: relative similarity which is the same with the above one and mutual information. It considers that the bigger these two thresholds are, the more interesting the pattern is. In these three thresholds, relative similarity indicates a linear relationship between genes while joint entropy (mutual information) measures a non-linear relationship. Taking them into consideration can guarantee that we can find potential relationships from the non-linear point of view based on the linear background.

We have conducted many experiments to identify our measures. The relationships between ncRNAs that we find result from the combination of similarity and entropy theory, which cannot be explored by other measures. We extract data from miRBase and RNAdb to construct three datasets: “let7”, “hsa” and “rna”, respectively. In “let7”, we analyze the relationships between miRNAs which belong to the same family but in different species. “hsa” is special for exploring the relationships between miRNAs all of which belong to *H.* sapiens but in different miRNA families. The “rna” set is composed of three ncRNA: miRNA, snoRNA and piRNA, which is used to find out the relationships between them. Based on these three datasets, we analyze the measures from two angles: biological significance and measure performance.

From the biological point of view, we first find interesting knowledge about species in dataset “let7”. All the patterns are identified by miRNAMap, which includes known miRNA information such as related gene list and gene target. Through querying, we find all seven patterns from “let7” can be identified in miRNAMap. We also find novel knowledge about species. For example, the relationship between *M. domestica* and *M. musculus* is closer than that between *M. domestica* and *G. gallus*, or that between *M. musculus* and *G. gallus*. In addition to this, our methods can reduce dramatically the number of related miRNAs for one special miRNA because we take entropy theory into account. For example our

measures find 19 miRNAs which are related with mdo-let-7i. However, there are 130 miRNAs related with it in miRNAMap. This means that we can find more interesting patterns with fewer results and faster speed, which obviously helps ones analyze the relationships between miRNAs. In "hsa", we extend our algorithms into cross-genome pattern analysis from cross-species. We find novel patterns and explain the reasons. These conclusions illustrate that our measures are useful to find novel relationships between ncRNAs and speculate the ncRNA functions.

We also found that our measures are suitable for any ncRNA dataset. They are not sensitive to the parameters and dataset types. That means the stable results can be achieved. Moreover, compared to joint entropy measure, we find that almost all patterns by mutual information measure can be divided into three categories. The advantage of this is that we can cluster the patterns easily and find out which part to focus on. The disadvantage of this is the sequences in our methods are inevitably limited by the current biological techniques. For example, the lengths of the hairpins in database miRBase are determined by the window length used in miRNA search programs. So in our methods, the sequence lengths we used might not be true. But this is a drawback that we cannot deal with currently. In addition, threshold setting is another issue that we will deal with.

All in all, the feature of our methods is that we focus on the relationships which are across genomes or species, which cannot be found by traditional sequence analysis tools or clustering. This relationship can contribute to explore the unknown genes' functions, the gene regulation mechanism. Further more, if we link all the relationships together, it can be considered to be a part of gene regulation network, which helps to understand how gene works when they regulate each other.

Acknowledgment

The work in this paper was partially supported by Australian Research Council Grant DP0559251.

References

[1] Xiao Z, Xue L, et al. Advances in the research technology of ncRNA. Chin Bull Life Sci 2007;2:122–5.
[2] Huang D et al. Identifying the biologically relevant gene categories based on gene expression and biological data: an example on prostate cancer. Bioinformatics 2007;12:1503–10.
[3] Yiu SM, Yiu SW, Lee LK, et al. Sharing and access right delegation for confidential documents: a practical solution. Inf Manage 2006;43:607–16.
[4] An J, Chen YPP. Finding rule groups to classify high dimensional gene expression datasets. Comput Biol Chem 2009;33:108–13.
[5] Cummins J, He Y, Leary RJ, et al. The colorectal microRNAome. PNAS 2006;10:3687–92.
[6] Joung J, Hwang KB, Nam JW, et al. Discovery of microRNA-mRNA modules via population-based probabilistic learning. Bioinformatics 2007;9:1141–7.
[7] Ng KL, Mishra SK, et al. De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures. Bioinformatics 2007;11:1321–30.
[8] Lagos QM, Rauhut R, Meyer J, et al. New microRNAs from mouse and human. RNA 2003;9:175–9.
[9] Altuvia Y, Landgraf P, Elefant N, et al. Clustering and conservation patterns of human microRNAs. Nucleic Acids Res 2005;8:2697–706.
[10] Maziere P, Enright AJ, et al. Prediction of microRNA targets. Drug Discov Today 2007;12:452–8.
[11] Zhang S, Chen F, Wu X, et al. Identifying bridging rules between conceptual clusters. In: Proceedings of 12th ACM SIGKDD international conference on knowledge discovery and data mining (KDD-06); 2006. p. 815–20.
[12] Gruber AR, Bernhart SH, Hofacker I, et al. Strategies for measuring evolutionary conservation of RNA secondary structure. BMC Bioinformatics 2008;9:122.
[13] Andronescu M, Bereg V, Hoos HH, et al. RNA STRAND: the RNA secondary structure and statistical analysis database. BMC Bioinformatics 2008;9:340.
[14] Moulton V. Tracking down noncoding RNAs. PNAS 2005;102:2269–70.
[15] Hsu WC et al. miRNAMap: genomic maps of microRNA genes and their target genes in mammalian genomes. Nucleic Acids Res. 2006;34:135–9.
[16] Hua Y, Huang HD, Hsu SD, et al. Progresses on the microRNA study. Chin Bull Life Sci 2005;3:1–4.
[17] Kong Y, Krichevsky A, Grad Y, et al. MicroRNA: biological and computational perspective. Genomics Proteomics Bioinformatics 2005;2:62–72.
[18] Chen Q, Chen YPP, Zhang C. Detecting inconsistency in biological molecular databases using ontology. Data Mining Knowledge Discov 2007;15:275–96.
[19] Lu J, Getz Gad, Miska EA. MicroRNA expression profiles classify human cancers. Nature 2005;9:834–8.
[20] Kim J et al. Identification of many microRNAs that copurify with polyribosomes in mammalian neurons. PNAS 2004;101:360–5.
[21] Huang Z, Chow TWS, et al. Genome-wide analyses of two families of snoRNA genes from *Drosophila melanogaster*, demonstrating the extensive utilization of introns for coding of snoRNAs. RNA 2005;11:1303–16.
[22] Cavaille J, Herve S, Paulsen M, et al. Identification of tandemly-repeated C/D snoRNA genes at the imprinted human 14q32 domain reminiscent of those at the Prader-Willi/Angelman syndrome region. Hum Mol Genet 2002;13:1527–38.
[23] Brown J, Clark GP, Leader DJ, et al. Multiple snoRNA gene clusters from *Arabidopsis*. RNA 2001;7:1817–32.
[24] Samarsky DA, Fournier MJ, Singer RH, et al. The snoRNA box C/D motif directs nucleolar targeting and also couples snoRNA synthesis and localization. EMBO J 1998;17:3747–57.
[25] Lau NC, Seto AG, Kim J, et al. Characterization of the piRNA complex from rat testes. Science 2006;313:363–7.
[26] Girard A, Sachidanandam R, Hannon GJ, et al. A germline-specific class of small RNAs binds mammalian Piwi proteins. Nature 2006;13:199–202.
[27] Houwing S, Kamminga LM, Berezikov E, et al. A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in zebrafish. Cell 2007;129:69–82.
[28] Alexei A, Ravi S, Angelique G, et al. Developmentally regulated piRNA clusters implicate MILI in transposon control. Science 2007;316:744–7.
[29] Bose Indranil. Deciding the financial health of dot-coms using rough sets. Inf Manage 2006;43:835–46.
[30] Mandayamvokkarne R, Giordana A, Yao FF, et al. On nearest-neighbor graph. Discrete Comput Geom 1997;17:263–82.
[31] Griffiths JS, Grocock RJ, Dongen SV, et al. MiRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res 2006;34:D140–4.
[32] Paddison PJ, Caudy AA, Bernstein E, et al. Short hairpin RNAs (shRNAs) induce sequence-specific silencing in mammalian cells. Genes Dev 2002;16:948–58.
[33] Zagrovic B et al. β-Hairpin folding simulations in atomistic detail using an implicit solvent model. J Mol Biol 2001;313:151–69.
[34] Zhang M, Zhou Y, Zhang W, et al. Apoptosis induced by short hairpin RNA-mediated STAT6 gene silencing in human colon cancer cells. Chin Med J 2006;119:801–8.
[35] Johnson S, Grosshans H, Shingara J, et al. RAS is regulated by the let-7 MicroRNA family. Cell 2005;120:635–47.
[36] Abbott AL, Alvarez SW, Miska EA, et al. The let-7 MicroRNA family members mir-48, mir-84, and mir-241 function together to regulate developmental timing in *Caenorhabditis elegans*. Dev Cell 2005;9:403–14.
[37] Li M, Jones-Rhoades MW, Lau NC, et al. Regulatory mutations of mir-48, a *C. elegans* let-7 family MicroRNA, cause developmental timing defects. Dev Cell 2005;9:415–22.
[38] Akao Y, Nakagawa Y, Naoe T, et al. let-7 MicroRNA functions as a potential growth suppressor in human colon cancer cells. Biol Pharm Bull 2006;29:903–6.
[39] Wulczyn F, Smirnova L, Rybak A, et al. Post-transcriptional regulation of the let-7 microRNA during neural cell specification. FASEB J Res Commun 2007;21:415–26.
[40] IBM Intelligent Information System. Available from: http://www.almaden.ibm.com/software/quest/resources/.
[41] Han J, Pei J, Yin Y, et al. Mining frequent patterns without candidate generation. Data Mining Knowledge Discov 2004;8:53–87.
[42] Washietl S, Hofacker IL, Stadler PF, et al. Fast and reliable prediction of noncoding RNAs. PNAS 2005;102:2454–9.
[43] Hillier L, Miller W, Birney E, et al. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 2004;7018:695–716.
[44] John B, Enright AJ, Aravin A, et al. Human MicroRNA targets. PLoS Biol 2004;11:1862–79.
[45] Lagos QM, Rauhut R, Yalcin A, et al. Identification of tissue-specific microRNAs from mouse. Curr Biol 2002;9:735–9.
[46] Michael MZ, O'Connor SM, Van NG, et al. Reduced accumulation of specific microRNAs in colorectal neoplasis. Mol Cancer Res 2003;12:882–91.
[47] Poy M, Eliasson L, Krutzfeldt J, et al. A pancreatic islet-specific microRNA regulate insulin secretion. Nature 2004;7014:226–30.
[48] Watanabe T, Takeda A, Tsukiyama T, et al. Identification and characterization of two novel classes of small RNAs in the mouse germline: retrotransposon-derived siRNAs in oocytes and germline small RNAs in testes. Genes Dev 2006;20:1732–43.
[49] Landgraf P, Rusu M, Sheridan R, et al. A mammalian microRNA expression atlas based on small RNA library sequencing. Cell 2007;7:1401–14.
[50] Lui WO, Pourmand N, Patterson BK, et al. Patterns of known and novel small RNAs in human cervical cancer. Cancer Res 2007;67:6031–43.
[51] Kasashima K, Nakamura Y, Kozu T, et al. Altered expression profiles of microRNAs during TPA-induced differentiation of HL-60 cells. Biochem Biophys Res Commun 2004;2:401–10.
[52] Chen YPP, Chen F. Targets for drug discovery using bioinformatics. Exp Opin Ther Targets 2008;12(4):383–9.

[53] An J, Chen YPP, Chen H. DDR: an index method for large time series datasets. Inf Syst 2005;30:333–48.
[54] Chen Q, Chen YPP. Mining frequent patterns for AMP-activated protein kinase regulation on skeletal muscle. BMC Bioinformatics 2006;7:394.
[55] Chen Q, Chen YPP. Discovery of structural and functional features in RNA pseudoknots. IEEE Trans Knowledge Data Eng 2009;21(7):974–84.
[56] Porkka KP, Pfeiffer MJ, Waltering KK, et al. MicroRNA expression profiling in prostate cancer. Cancer Res 2007;67:6130–5.
[57] Tran TT, Zhou F, Marshburn S, et al. De novo computational prediction of non-coding RNA genes in prokaryotic genomes. Bioinformatics 2009;25:2897–905.
[58] Kavanaugh LA, Dietrich FS. Non-coding RNA prediction and verification in *Saccharomyces cerevisiae*. PLoS Genet 2009;5:e1000321.
[59] Karklin Y, Meraz RF, Holbrook SR. Classification of non-coding RNA using graph representations of secondary structure. In: Pacific symposium on biocomputing, vol. 10; 2005. p. 4–15.
[60] Weinberg Z, Ruzzo WL. Faster genome annotation of non-coding RNA families without loss of accuracy. In: Proceedings of the eighth annual international conference on research in computational molecular biology; 2004. p. 243–51.
[61] Klein RJ, Misulovin Z, Eddy SR. Noncoding RNA genes identified in AT-rich hyperthermophiles. PNAS 2002;99:7542–7.
[62] Lindgreen S, Gardner PP, Krogh A. Measuring covariation in RNA alignments: physical realism improves information measures. Bioinformations 2006;22:2988–95.
[63] Freyhult E, Gardner P, Moulton V. A comparison of RNA folding measures. BMC Bioinformatics 2005;6:241.
[64] Huynen M, Gutell R, Konings D. Assessing the reliability of RNA folding using statistical mechanics. J Mol Biol 1997;267:1104–12.
[65] Rodland EA. Pseudoknots in RNA secondary structures: representation, enumeration and prevalence. J Comput Biol 2006;13:1197–213.